— SMGloM Blue Note[*] —
SMGloM: a Semantic Multilingual Glossary System for Mathematics

Michael Kohlhase
Computer Science, Jacobs University Bremen
http://kwarc.info/kohlhase

July 22, 2014

Abstract

Mathematical vernacular – the everyday language we use to communicate about mathematics is characterized by a special vocabulary. If we want to support humans with mathematical documents, we need a resource that captures the terminological, linguistic, and ontological aspects of the mathematical vocabulary. In the SMGloM project and system, we aim to do just this. In this note we present the glossary system prototype, the content organization, and the envisioned community aspects.

Contents

[*] Inspired by the "blue book" in Alan Bundy's group at the University of Edinburgh, SMGloM blue notes, are documents used for fixing and discussing ϵ-baked ideas in projects by the SMGloM group (see http://mathhub.info/help/SMGloM). Unless specified otherwise, they are for project-internal discussions only. Please only distribute outside the SMGloM group after consultation with the author.

1 Introduction

Mathematics plays a fundamental role in science, technology, and engineering (STEM). Mathematical knowledge is rich in content, sophisticated in structure, and technical in presentation. Its conservation, dissemination, and utilization constitutes a challenge for the community and an attractive line of inquiry.

One of the challenging aspects of mathematical language is its special terminology of technical terms that are defined in various mathematical documents. To alleviate this, mathematicians use special glossaries, traditionally lists of terms in a particular domain of knowledge with the definitions for those terms. Originally, glossaries appeared as alphabetical lists of new/introduced terms with short definitions in the back of books to help readers understand the contents.

On the web we find many examples of "mathematical glossaries" for middle and highschool mathematics, ranging from [MCG], which contains ca. 150 glossary entries (technical terms with their definitions) to [HRW], which has ca 500 terms in 13 languages; the British Sign Language Math Glossary [BSL] introduces the signs for ca. 50 mathematical concepts via short video as part of a resource that helps teach science to the hearing impaired.

For university-level and research mathematics, similar resources exist, but they usually call themselves encyclopedias. They range from PlanetMath.org [PM], with 9,000 "articles", over the Encyclopedia of Mathematics [EM] with more than 8,000 entries, illuminating nearly 50,000 notions in graduate-level mathematics and Eric Weisstein's Mathworld [Wei] to the wikipedia which contains a large number of pages on mathematical topics – [MG11] estimate 28,000. All of these resources are organized lexically, but transcend the glossary format by complementing the definition with historical remarks, accounts of salient properties, and relations to other objects.

Another kind of resource that deals with terminology of mathematics are "dictionaries", which align mathematical terms in different languages by their meaning (originally without) giving a definition. Examples include [ES82] is a printed English/German/French/Russian dictionary with about 35,000 lexical entries and the public domain [CMD] an online German/English/Bulgarian dictionary for mathematics.

In the last decades the term "glossary" has also been applied to digital vocabularies (online encyclopedias, thesauri, dictionaries, etc.), which have become important resources in knowledge-based systems. This is especially true for vocabularies that have a *i*) semantic aspect – i.e. some of the relations are made explicit and machine-actionable, they are also called "ontologies" – or *ii*) that are multilingual. Digital vocabularies can be hand-curated, or machine-generated/collected; an example of the former is the WordNet lexical database for English [Fel98; WN], an example of the latter is DBPedia [Aue+07; DBP13], but they can also be hybrid, e.g. the UWN/Menta project [MW09; YAGO] generates a multilingual WordNet by automatically adding other languages by crawling Wikipedia.

In this note we present the SMGloM project, which aims to create a semantic, multilingual glossary for mathematics. This resource combines the characteristics of dictionaries and glossaries, with those of ontologies, but restricts the content to definitions and the relations to the lexical ones to keep the task manageable. Here we give a high-level overview over the data model, the SMGloM system, organizational and legal issues, possible applications, and the state of the effort of seeding the glossary. Details can be found in additional blue notes [**Kohlhase:edsmglom14**; Koh14a; Koh14b].

2 Data Model and Encoding

We build the data model of SMGloM on top of the one of OMDoc/Mmt from [Koh06; RK13], which provides views, statements, and theories; see [Koh+13; Koh14c] for more accessible descriptions, which we assume as background references.

In a nutshell (details can be found in [Koh14a]), a **glossary entry** consists of one **symbol**, its **definition**, and a set of **verbalizations** and **notations**. A symbol is a formal identifier of a mathematical object/concept (i.e a formal object). The verbalizations relate it to lexical entries (identified by the stem of the head), which we call **glossary terms**[1] EdN:1

The definitions could be written down in a formal logic, but in the SMGloM, we write them down in mathematical vernacular (common mathematical language; in SMGloM natural language with sTeX annotations). Thus we consider "the definition" of a symbol to be given by a set of vernacular definitions, which are assumed to be translations of each other – an important structural invariant of the SMGloM that needs to be maintained.

Glossary entries are often grouped into a **glossary module**, which is represented as $n+1$ OMDoc/Mmt theories: one for the language-independent part (called the **module signature**, it introduces the symbols, their dependencies, and notations), and n for the **language bindings** (which introduce the definitions and verbalizations of symbols).

3 Organizing a Communal Resource

The ultimate cause of the SMGloM project and system is to facilitate the establishment of a knowledge resource for mathematics. We need to take appropriate organizational measures to support this.

We are currently establishing a wiki-like archive submission system for glossary modules on MathHub [MH] and thinking of a quality assurance system that is based on a community/karma-driven approval system. Openness and semantic stability are ensured by a special licensing and publication regime: The SMGloM license protects symbols against non-conservative changes while allowing derived works; see [Koh14b] for details.

4 A System and Home for SMGloM

Naturally, a complex resource like the SMGloM has to be supported by a system that partially automates editing, management, refactoring, quality control, etc. Instead of building a system from scratch, we make use of MathHub [MH], a portal for active mathematical documents and an archive for flexiformal mathematics, and extend it with SMGloM-specific functionality.[2] EdN:2

4.1 SMGloM Editing Workflows

[**Kohlhase:edsmglom14**]

[1] EdNote: MK: I am not clear whether we should also give the notations a similar status of "glossary terms"; if so, we have to invent the concept.

[2] EdNote: continue, and describe this better in a blue note.

branch

1. Module `graph`
   syno- hyper- hypo- mero-nyms de ro zh
   A **graph** is a structure $\langle V, E\rangle$ such that V is a set and $E \subseteq V \times V$ is a subset of the set of pairs from V. We call V the vertices (or **node**s, **point**s, **junction**s) and E the **edge**s (or **line**s, **branch**es, **arc**s) of G.
2. Module `inverse function`
   syno- hyper- hypo- mero-nyms de ro zh
   A **branch** of a multivalued function f is a univalent sub-relation $b \subseteq f$.
3. ...

branch curve

Figure 1: Sample glossary interface for SMGloM

5 Applications of the SMGloM

The main advantage of SMGloM over existing terminological resources for mathematics is that it makes important linguistic and ontological relations explicit that these do not. This extension makes a large variety of applications feasible without requiring full formalization, the cost of which would be prohibitive. We will sketch some of the applications here.

5.1 Glossary of Mathematical Terms

An interface that presents SMGloM like a traditional glossary, i.e. as a (sorted) list of glossary entries. In addition, the semantic information in SMGloM can be used to adequately mark up references to as well as relations with (e.g. "synonym of", or "translation of") other entries. See Figure 1 for an example. There can be sub-glossaries, for certain areas of mathematics, for certain languages, etc.

5.2 Flexible Styling/Presentation

If we have formulae in content markup (i.e. in content MathML e.g. in OMDoc or sTeX), then we can adapt the rendering of formulae with symbols that having multiple notations in SMGloM to the user's preferences. Then, each user can state their notational preferences (in terms of SMGloM notation definitions), and the formulae in SMGloM will be rendered using these, adapting to the preferences of the reader.

5.3 Notation-Based-Parsing

The notation definitions from SMGloM can be seen as user-contributed grammar rules. Therefore, they can be used for parsing formulae from presentation to content markup in the longer run. This will lead to a context-sensitive formula parser, where "context" is defined by the SMGloM glossary modules currently in focus – here the data model in term of OMDoc/Mmt theories directly contributes to the applications of the SMGloM. See also [Koh13] for details.

5.4 More Semantic Search

As SMGloM declares symbols together with notations, definitions and verbalizations it provides an unique opportunity for applying semantic search services based on it in a variety of settings:

1. notation-based parsing in the input phase could make formula entry into an interactive disambiguation process. For instance, a user enters e^?x, and the system ask her: "with e, do you mean Euler's number?", and also: "Is $e^?x$ a power operation?". The answers will then help refine the search.
2. Alternatively, search could use disambiguation as a facet in the search to refine the results or for clustering the results.
3. Furthermore, the SMGloM information could be used for query expansion (both visible or automatic): if the user searches for e, then the query could be expanded e.g. by
   - the string Euler's Number (there is an interesting question about what to do with the language dependency here) and even
   - the formula $\lim_{?n\to\infty}(1+\frac{1}{?n})^{?n}$ ($?n$ is a query variable).

   Other query expansions based on SMGloM knowledge should be looked at.

5.5 Verbalization-Based Translation

One of the most tedious parts of translating mathematical documents is the correct use of technical terms. A semantically preloaded text (i.e. one that has all formulae in content markup and many semantic objects explicitly marked up) can be term-translated automatically using the translation relation induced by SMGloM. Of course, synonyms must be resolved consistently (there has to be an interface for this). This (and related semantic tasks) are for domain specialists. The intervening text can be done by lesser trained individuals (or even a variant of google translate). This will make translations much cheaper and will make math available in more languages.

A particularly interesting avenue of development is to integrate the glossary with an CAT (Computer Aided Translation) tool like the open-source OmegaT [OTa; OTb]. On the one hand, information from the glossary can be exported for use in OmegaT translations by exporting into their terminology database. On the other hand OmegaT could be used for translating SMGloM entries. This seems doable, since OmegaT has a LaTeX mode.

5.6 Wikifiers like NNexus

Wikifiers are systems that given a glossary of terms create definitional links in documents. A math-specific example is the NNexus system [**GinCor:nnexusCICM13**], it can already use the SMGloM glossary.

6 Seeding the SMGloM

To make public contributions to SMGloM feasible, it must already contain a nucleus of (basic) entries that can be referenced in other glossary components. The SMGloM project is currently working towards a basic inventory of glossary entries, and has almost arrived at the first milestone of 500 entries – most with two language bindings. The current glossary contains

- ca. 150 glossary entries from elementary mathematics, to provide a basis for further development
- ca. 350 are special concepts from number theory to explore the suitability of the SMGloM for more advanced areas of mathematics
- a handfull views to establish the concept and serve as examples.

Once the authoring and quality assurance systems have progressed sufficiently (ETA summer 2014), we want to open the system to the mathematical community. By that time, we hope to be able to show some of the applications sketched in Section 5.

We should identify existing open glossaries and incorporate them into the SMGloM format (semi)-automatically. One source is to try to identify definitions from open online[1] encyclopedia like PlanetMath.org or Wikipedia, the latter could even be a source for multilingual entries. We can also directly glean "proposed" glossary entries and their bindings from [CMD] an online German/English/Bulgarian dictionary for mathematics; it is in the public domain.

7 Open Questions

Here we tabulate some of the open questions as they come along.

7.1 Adjectives

Adjectives are interesting structures both grammatically as well as in their use as qualifiers in maths. Consider for instance the case of "Boolean": a "Boolean Ring" is Boolean in a very different way than a "Boolean Function" – the first is about idempotency of multiplication and the the second describes the type of arguments and results. But the adjective "Abelian" applies to many structures and basically says that the main binary operation is commutative. So whereas we almost certainly want separate glossary entries for the "Boolean Ring" and "Boolean Function", we might want to have a single entry for "Abelian".

Weak type theory [KN04] and the Mizar language [Miz] have intuitions about this, which could help in the SMGloM system and data model.

8 Conclusion

We have presented first ideas for the data model a semantic, multilingual glossary for mathematics and sketched an sTeX-based encoding as well as a Planetary-based knowledge portal. This note is mainly intended to collect feedback.

The implementation work on the sTeX bindings (for OMDoc translation) and the SMGloM system is ongoing, and in an early state.[3] EdN:3

References

[Aue+07] Sören Auer et al. "DBpedia: A Nucleus for a Web of Open Data". In: *The Semantic Web*. 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007. (Busan, Korea, Nov. 11–15, 2007). Ed. by Karl Aberer et al. LNCS 4825. Springer Verlag, 2007.

[BSL] *British Sign Language Math Glossary*. URL: `http://www.ssc.education.ed.ac.uk/bsl/maths.html` (visited on 02/01/2014).

[CMD] *Math Dictionary*. URL: `http://mathdict.chitanka.info/` (visited on 02/01/2014).

[DBP13] *DBpedia*. Sept. 17, 2013. URL: `http://dbpedia.org` (visited on 02/21/2014).

[1] Unfortunately Wolfram MathWorld is closed.

[3] EdNote: write about availability and demo installation.

[EM] *Encyclopedia of Mathematics*. URL: http://www.encyclopediaofmath.org (visited on 02/01/2014).

[ES82] Günther Eisenreich and Ralf Sube. *Wörterbuch Mathematik Englisch Deutsch Französisch Russisch*. Verlag Harri Deutsch, 1982.

[Fel98] Christiane Fellbaum, ed. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[HRW] *HRW: Middle School Math – Multilingual Glossary*. URL: http://my.hrw.com/math06_07/nsmedia/tools/glossary/msm/glossary.html (visited on 02/01/2014).

[KN04] Fairouz Kamareddine and Rob Nederpelt. "A refinement of de Bruijn's formal language of mathematics". In: *Logic, Language and Information* 13.3 (2004), pp. 287–340.

[Koh+13] Michael Kohlhase et al. "Zentralblatt Column: Mathematical Formula Search". In: *EMS Newsletter* (Sept. 2013), pp. 56–57. URL: http://www.ems-ph.org/journals/newsletter/pdf/2013-09-89.pdf.

[Koh06] Michael Kohlhase. OMDOC – *An open markup format for mathematical documents [Version 1.2]*. LNAI 4180. Springer Verlag, Aug. 2006. URL: http://omdoc.org/pubs/omdoc1.2.pdf.

[Koh13] Michael Kohlhase. "Formula Understanding with Notation Definitions". LaMaPUn Blue Note. 2013. URL: https://svn.kwarc.info/repos/lamapun/doc/blue/ndparse/note.pdf.

[Koh14a] Michael Kohlhase. "A Data Model and Encoding for SMGloM". SMGloM Blue Note. 2014. URL: http://gl.mathhub.info/smglom/smglom-doc/raw/master/source/blue/datamdl/note.pdf.

[Koh14b] Michael Kohlhase. "Content Management in SMGloM". SMGloM Blue Note. 2014. URL: http://gl.mathhub.info/smglom/smglom-doc/raw/master/source/blue/contmgt/note.pdf.

[Koh14c] Michael Kohlhase. "Mathematical Knowledge Management: Transcending the One-Brain-Barrier with Theory Graphs". In: *EMS Newsletter* (2014). in press. URL: http://kwarc.info/kohlhase/papers/ems14.pdf.

[MCG] *Math.com Glossary*. URL: http://www.math.com/school/glossary/glossindex.html (visited on 02/01/2014).

[MG11] Jozef Misutka and Leo Galambos. "System Description: EgoMath2 As a Tool for Mathematical Searching on Wikipedia.org". In: *Calculemus/MKM*. Ed. by James Davenport et al. LNAI 6824. Springer Verlag, 2011, pp. 307–309.

[MH] *MathHub.info: Active Mathematics*. URL: http://mathhub.info (visited on 01/28/2014).

[Miz] *Mizar*. URL: http://www.mizar.org (visited on 02/27/2013).

[MW09] Gerard de Melo and Gerhard Weikum. "Towards a Universal Wordnet by Learning from Combined Evidence". In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*. Hong Kong, China: ACM, 2009, pp. 513–522. DOI: http://doi.acm.org/10.1145/1645953.1646020.

[OTa] *OmegaT*. URL: `http://omegat.org/` (visited on 02/04/2014).

[OTb] *OmegaT – multiplatform CAT tool*. URL: `http://sourceforge.net/projects/omegat/` (visited on 02/04/2014).

[PM] *PlanetMath.org – Math for the people, by the people*. URL: `http://planetmath.org` (visited on 11/11/2012).

[RK13] Florian Rabe and Michael Kohlhase. In: *Information & Computation* 230 (2013), pp. 1–54. URL: `http://kwarc.info/frabe/Research/mmt.pdf`.

[Wei] Eric W. Weisstein, ed. *Wolfram MathWorld. the web's most extensive mathematics resource*. Wolfram Research. URL: `http://mathworld.wolfram.com` (visited on 12/02/2009).

[WN] *WordNet: A lexical database for English*. URL: `https://wordnet.princeton.edu/` (visited on 05/26/2013).

[YAGO] *Towards a Universal Multilingual Wordnet*. URL: `http://www.mpi-inf.mpg.de/yago-naga/uwn/` (visited on 05/26/2013).